Maximum Likelihood Sound Source Localization and Beamforming for Directional Microphone Arrays in Distributed Meetings

Cha Zhang, *Member, IEEE*, Dinei Florêncio, *Senior Member, IEEE*, Demba E. Ba, and Zhengyou Zhang, *Fellow, IEEE*

***Abstract*—In distributed meeting applications, microphone arrays have been widely used to capture superior speech sound and perform speaker localization through sound source localization (SSL) and beamforming. This paper presents a unified maximum likelihood framework of these two techniques, and demonstrates how such a framework can be adapted to create efficient SSL and beamforming algorithms for reverberant rooms and unknown directional patterns of microphones. The proposed method is closely related to steered response power-based algorithms, which are known to work extremely well in real-world environments. We demonstrate the effectiveness of the proposed method on challenging synthetic and real-world datasets, including over six hours of recorded meetings.**

***Index Terms*—Beamforming, directional mics, microphone array, sound source localization.**

I. Introduction

ELECTRONICALLY steerable arrays of microphones have recently found a variety of new applications, such as human-computer interaction [1], [2], and intelligent rooms [3]–[5]. A microphone array-based system has a number of advantages over a single microphone system. For instance, it may be electronically aimed to capture an audio signal from a desired source location and simultaneously attenuate environmental noises. It can also be used to localize an active speaker nearby, allowing computer controlled devices to provide a speaker location-aware user interface. To give a more concrete example, a distributed meeting device called RoundTable [6] is shown in Fig. 1(a). It has a six-element circular microphone array at the base, and five video cameras at the top. The captured videos are stitched into a 360-degree panorama, which gives a global view of the meeting room (but at low resolution due to bandwidth constraints). The microphone array is there not only to capture superior sounds, but also to detect the sound source location and generate a high-resolution video of the speaker for a better viewing experience. The device enables remote group members to hear and view meetings live online. In addition, the meetings can be recorded and archived, allowing people to browse them afterwards.

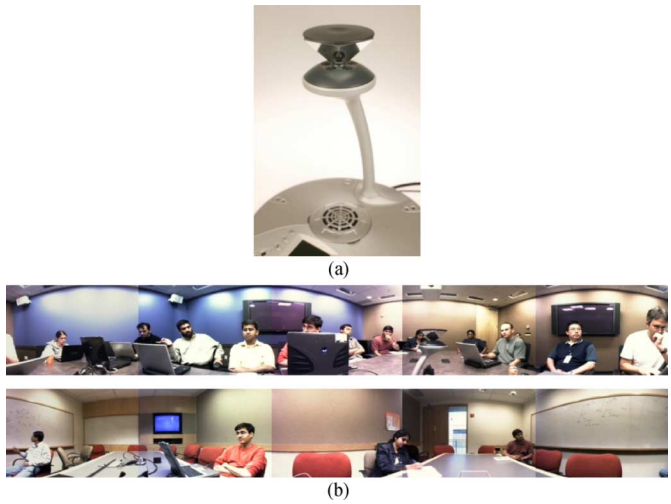


Fig. 1. RoundTable and its captured images. (a) RoundTable device. (b) Two images captured by the device.

The key technologies involved in microphone arrays are *sound source localization (SSL)* and *beamforming*, both having been active research topics since the 1970s [7]–[9]. A large number of SSL and beamforming algorithms have been proposed in the literature, of varying degrees of accuracy and computational complexity. In this paper, we limit our attention to algorithms that are applicable in distributed meeting scenarios, such as the set of microphones in the RoundTable device. These microphone arrays often bear the following set of characteristics.

- The number of microphones in a single device is often limited, e.g., four, six, or eight. Linear arrays and circular arrays are the most popular ones.
- Both omnidirectional and directional microphones are popular. In the case of the circular array in RoundTable, directional microphones are preferred due to their superior sound capture capability.
- Microphones in distributed meeting devices tend to be spaced at distances of around 10 cm. The difference in source energy at the microphone locations is not significant.

Manuscript received February 16, 2007; revised December 2, 2007. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Cha Zhang.

C. Zhang, D. Florêncio, and Z. Zhang are with Microsoft Research, Redmond, WA 98052 USA (e-mail: chazhang@microsoft.com; dinei@microsoft.com; zhang@microsoft.com).

D. E. Ba is with the Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology, Cambridge, MA 02139 USA.

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TMM.2008.917406

- Meeting rooms can receive special sound-treatment, or not. In the latter case, reverberation could be significant due to common reflective objects such as whiteboards or displays.
- Many distributed meeting devices need to function properly without a linked computer. The available computational resources are thus very limited.

Given these characteristics, the choices for SSL and beamforming algorithms are very limited. For instance, SSL algorithms that rely on sensing the difference in source energy among different microphones cannot be applied due to the close distance between microphones. If the distributed meeting device is to be mass produced, measuring the microphones' directional response patterns for each device would be extremely difficult if not impossible, hence the algorithm must adapt to microphones with unknown gains. The algorithm also has to be very robust to reverberation, which could change significantly from room to room in the real world. Lastly, any algorithm used in such devices has to be computationally very efficient.

In this paper, we present a maximum likelihood (ML) framework for microphone array sound source localization and beamforming. While this is not the first time ML estimation is applied for SSL or beamforming [10]–[13], this paper builds a much stronger connection between the proposed ML-based SSL (ML-SSL) and the popular steered response power (SRP) based algorithms, which are known to work extremely well in practical environments [3], [14], [15] and have very low computational cost. We demonstrate within the ML framework how reverberation can be dealt with by introducing an additional term during noise modeling, and how the unknown directional patterns of microphone gains can be compensated for from the received signal and the noise model. The result is a new and efficient SSL algorithm that can be applied to various kinds of microphone arrays, even for challenging cases such as circular directional arrays with unknown directional patterns (e.g., the array in RoundTable). The effectiveness of the proposed method is shown on both synthetic and real-world data. The synthetic data allows a more precise study of the influence of noise level and reverberation in the algorithm performance. The extensive real-world data corroborates the improvement in relevant scenarios. This data consists of 99 sequences, totaling over six hours of meetings, recorded in over a dozen different meeting rooms.

Additionally, our ML derivation demonstrates that the traditional minimum variance distortionless response (MVDR) beamforming technique is equivalent to the ML-SSL. In other words, we show that the result of ML-SSL is the same as if one used multiple MVDR beamformers to perform beamforming along multiple hypothesis directions and picked the output direction which results in the highest signal-to-noise ratio (SNR). The technique proposed above to handle unknown directional patterns of microphone gains can thus be extended to MVDR. We call the revised algorithm enhanced MVDR (eMVDR), and show that it outperforms the traditional method for circular directional microphone arrays.

The rest of the paper is organized as follows. We review a number of related SSL and beamforming approaches in Section II. The ML framework is derived in Section III. Using the proposed framework, we derive an efficient SSL algorithm and compare it with various existing approaches in Section IV. eMVDR is discussed in Section V. Experimental results and conclusions are given in Sections VI and VII, respectively.

II. Review of Existing Approaches

We now review some existing SSL and beamforming approaches that are closely related to the proposed algorithm.

A. SSL

For broadband acoustic source localization applications, such as teleconferencing, a number of SSL techniques are popular, including those based on the steered-beamformer (SB), high-resolution spectral estimation, time delay of arrival (TDOA) [9], and learning [16]. Among them, the TDOA-based approaches have received extensive investigation [3], [9], [17]–[20].

Consider an array of P microphones. Given a source signal $s(t)$, the signals received at these microphones can be modeled as [7], [9], [18], [20]

$$x_i(t) = \alpha_i s(t - \tau_i) + n_i(t) \tag{1}$$

where $i = 1, \ldots, P$ is the index of each microphone; τ_i is the propagation delay from the source location to the i^{th} microphone location; α_i is a gain factor (including the effects of the propagation energy decay, the gain of the corresponding microphone, the directionality of the source and the microphone, etc.), and $n_i(t)$ is the noise sensed by the i^{th} microphone. Depending on the application, this noise term could include a room reverberation term to increase the robustness of the derived algorithm [15], [19], which will be discussed in detail in Section IV.

It is usually more efficient to work in the frequency domain, where we can rewrite the above model as

$$X_i(\omega) = \alpha_i(\omega) S(\omega) e^{-j\omega\tau_i} + N_i(\omega). \tag{2}$$

We can rewrite the above equation into a vector form as

$$\mathbf{X}(\omega) = S(\omega)\mathbf{G}(\omega) + \mathbf{N}(\omega), \tag{3}$$

where

$$\begin{aligned}\mathbf{X}(\omega) &= [X_1(\omega), \ldots, X_P(\omega)]^T \\ \mathbf{G}(\omega) &= [\alpha_1(\omega)e^{-j\omega\tau_1}, \ldots, \alpha_P(\omega)e^{-j\omega\tau_P}]^T \\ \mathbf{N}(\omega) &= [N_1(\omega), \ldots, N_P(\omega)]^T.\end{aligned}$$

Among the variables, $\mathbf{X}(\omega)$ represents the received signals, hence it is known. $\mathbf{G}(\omega)$ can be estimated or hypothesized during the computation process, which will be detailed later.

The most straightforward SSL algorithm involves taking a pair of microphones and estimating the difference in time of arrival by finding the peak of the cross-correlation (the direction of arrival is obtained by a geometric transformation from the time of arrival difference and the distance of the mics). For instance, the correlation between the signals received at microphone i and k is

$$R_{ik}(\tau) = \int x_i(t) x_k(t - \tau) dt. \tag{4}$$

The τ that maximizes the above correlation is the estimated time delay between the two signals. In practice, the above cross-correlation function can be computed more efficiently in the frequency domain as

$$R_{ik}(\tau) = \int X_i(\omega) X_k^*(\omega) e^{j\omega\tau} d\omega \tag{5}$$

where * represents the complex conjugate. If we substitute (2) into (5), and assuming noise and source signals are independent, the τ that maximizes the above correlation is $\tau_i - \tau_k$, which is the actual delay between the two microphones.

To find the τ that maximizes (5), one simple and extendable solution is through hypothesis testing. That is, hypothesize the source at certain location $\mathbf{s}$, which can be used to compute τ_i and τ_k. The hypothesis that achieves the highest cross-correlation is the resultant source location. When more than two microphones are considered, we sum over all possible pairs of microphones (including self pairs) and have

$$\begin{aligned}\mathcal{R}(\mathbf{s}) &= \sum_{i=1}^{P}\sum_{k=1}^{P} R_{ik}(\tau_i - \tau_k) \\ &= \sum_{i=1}^{P}\sum_{k=1}^{P} \int X_i(\omega) X_k^*(\omega) e^{j\omega(\tau_i-\tau_k)} d\omega \qquad (6) \\ &= \int \left[\sum_{i=1}^{P} X_i(\omega) e^{j\omega\tau_i}\right]\left[\sum_{k=1}^{P} X_k(\omega) e^{j\omega\tau_k}\right]^* d\omega \\ &= \int \left|\sum_{i=1}^{P} X_i(\omega) e^{j\omega\tau_i}\right|^2 d\omega. \qquad (7)\end{aligned}$$

Again we can solve the maximization problem through hypothesis testing on potential source locations $\mathbf{s}$. Equation (7) is also known as the steered response power (SRP) of the microphone array.

To address the reverberation and noise that may affect SSL accuracy, researchers have found that adding a weighting function in front of the correlation can greatly help. Equation (6) is thus rewritten as

$$\mathcal{R}(\mathbf{s}) = \sum_{i=1}^{P}\sum_{k=1}^{P} \int \Psi_{ik}(\omega) X_i(\omega) X_k^*(\omega) e^{j\omega(\tau_i-\tau_k)} d\omega \tag{8}$$

where $\Psi_{ik}(\omega)$ is a weighting function. A number of weighting functions have been investigated in the literature [7]. Among them, the heuristic-based PHAT weighting is defined as

$$\Psi_{ik}(\omega) = \frac{1}{|X_i(\omega)X_k^*(\omega)|} = \frac{1}{|X_i(\omega)||X_k(\omega)|}. \tag{9}$$

PHAT has been found to perform very well under realistic acoustic conditions [3], [15]. Inserting (9) into (8), we get

$$\mathcal{R}(\mathbf{s}) = \int \left|\sum_{i=1}^{P} \frac{X_i(\omega) e^{j\omega\tau_i}}{|X_i(\omega)|}\right|^2 d\omega. \tag{10}$$

This algorithm is called SRP-PHAT [14]. Note that SRP-PHAT is very efficient to compute, because the time complexity drops from P^2 in (8) to P.

A more theoretically-sound weighting function is the ML formulation derived by Brandstein *et al.* [9] under an assumption of high SNR and no reverberation. The weighting function of a microphone pair is defined as

$$\Psi_{ij}(\omega) = \frac{|X_i(\omega)||X_j(\omega)|}{|N_i(\omega)|^2|X_j(\omega)|^2 + |N_j(\omega)|^2|X_i(\omega)|^2}. \tag{11}$$

Equation (11) can be inserted into (8) to obtain an ML-based algorithm. This algorithm is known to be robust to noises, but its performance in real-world applications is relatively poor, because reverberation is not modeled in its derivation. In [15], Rui and Florêncio developed an improved version by considering the reverberation explicitly in the noise term. In a manner similar to the formulation in [3], the reverberation is treated in [15] as another type of noise, i.e.,

$$|N_i^c(\omega)|^2 = \gamma|X_i(\omega)|^2 + (1-\gamma)|N_i(\omega)|^2 \tag{12}$$

where $N_i^c(\omega)$ is now the combined noise or total noise. The first term on the right side of (12) is based on the assumption that the reverberation noise energy is proportional to the source signal energy. Equation (12) is then substituted into (11) (replacing $N_i(\omega)$ with $N_i^c(\omega)$) to obtain the new weighting function. Follow-up work [21] proposed a further approximation to yield

$$\mathcal{R}(\mathbf{s}) = \int \left|\sum_{i=1}^{P} \frac{X_i(\omega) e^{j\omega\tau_i}}{\gamma|X_i(\omega)| + (1-\gamma)|N_i(\omega)|}\right|^2 d\omega \tag{13}$$

whose computational efficiency is close to that of SRP-PHAT.

Note, however, that algorithms derived from (11) are not true ML algorithms for multiple microphones. This is because the optimal weight in (11) was derived only for two microphones. When more than two microphones are used, the adoption of (8) assumes that pairs of microphones are independent and hence that their likelihoods can be multiplied together, which is questionable. In this paper, a true ML algorithm will be developed for the case of multiple microphones. We will show the connection between the new algorithm and the existing algorithms in Section IV.

B. Beamforming

Beamforming refers to the technique that aims at improving captured sound quality by exploiting the diversity in the received signals of the microphone array. Depending on the location of the source and the interference, beamforming sets different gains to each mic to achieve its goal of noise suppression. Early designs were generally "fixed" beamformers (e.g., delay-and-sum), adapting only to the location of the desired source. More recent designs are based on "null-steering", and adapt to characteristics of the interference as well. The minimum variance distortionless response (MVDR) beamformer and its associated adaptive algorithm, the generalized sidelobe canceler (GSC) [22], [23], are probably the most widely studied and used beamforming algorithms, and form the basis of some commercially available arrays [24].

Assuming the direction of arrival (DOA) of the desired signal is known, we would like to determine a set of weights $\mathbf{w}(\omega)$,

such that $\mathbf{w}^H(\omega)\mathbf{X}(\omega)$ is a good estimate of $S(\omega)$. Note $\mathbf{X}(\omega)$ and $S(\omega)$ were defined in (3); the superscript H represents the Hermitian transpose. The beamformer that results from minimizing the variance of the noise component of $\mathbf{w}(\omega)^H\mathbf{X}(\omega)$, subject to a constraint of unity gain in the DOA direction, is known as the MVDR beamformer. The corresponding weight vector $\mathbf{w}(\omega)$ is the solution to the following optimization problem:

$$\min_{\mathbf{w}(\omega)} \mathbf{w}(\omega)^H\mathbf{Q}(\omega)\mathbf{w}(\omega), \text{ s.t. } \mathbf{w}(\omega)^H\mathbf{G}(\omega) = 1 \tag{14}$$

where $\mathbf{Q}(\omega)$ is the covariance matrix of the noise component:

$$\mathbf{Q}(\omega) = E[\mathbf{N}(\omega)\mathbf{N}^H(\omega)]. \tag{15}$$

In general, $\mathbf{Q}(\omega)$ is estimated from the data and therefore inherently contains information about the location of the sources of interference, as well as the effect of the sensors on those sources.

The optimization problem in (14) has an elegant closed-form solution [25] given by

$$\mathbf{w}(\omega) = \frac{\mathbf{Q}(\omega)^{-1}\mathbf{G}(\omega)}{\mathbf{G}(\omega)^H\mathbf{Q}(\omega)^{-1}\mathbf{G}(\omega)}. \tag{16}$$

Note that the denominator of (16) is merely a normalization factor which enforces the unity gain constraint in the look direction.

In practice, the DOA of the desired signal is not known exactly, which significantly degrades the performance of the MVDR beamformer [26]. Significant effort has gone into a class of algorithms known as robust MVDR [25], [27]. As a general rule, these algorithms work by specifying a region instead of a single look direction where the source has near-unity gain. Little attention has been paid to the gain term $\mathbf{G}(\omega)$ in (16), and most existing work assumes that it is either known or that the $\alpha(\omega)$ term can be ignored. This works well for linear arrays of omni-directional or directional microphones, where the gains of the microphones in the same direction are similar. However, for the circular geometry such as that of RoundTable, this directionality is accentuated: each microphone will have a significantly different direction of arrival in relation to the desired source. In this paper, we will address this issue by estimating the $\alpha(\omega)$ term explicitly during the beamforming process.

III. The Maximum Likelihood Framework

To assure a mathematically tractable solution, we assume the noise of the microphones follows a zero-mean, independent between frequencies, joint Gaussian distribution, i.e.,

$$p_\omega(\mathbf{N}(\omega)) = \rho_\omega \exp\left\{-\frac{1}{2}[\mathbf{N}(\omega)]^H\mathbf{Q}^{-1}(\omega)\mathbf{N}(\omega)\right\} \tag{17}$$

where ρ_ω is a normalization constant. When the covariance matrix $\mathbf{Q}(\omega)$ can be calculated/estimated from known signals, the likelihood of the received signals can be written as

$$p(\mathbf{X}|S,\mathbf{G},\mathbf{Q}) = \prod_\omega p_\omega(\mathbf{X}(\omega)|S(\omega),\mathbf{G}(\omega),\mathbf{Q}(\omega)) \tag{18}$$

where

$$p_\omega(\mathbf{X}(\omega)|S(\omega),\mathbf{G}(\omega),\mathbf{Q}(\omega)) = \rho_\omega \exp\left\{-\frac{J(\omega)}{2}\right\} \tag{19}$$

and

$$J(\omega) = [\mathbf{X}(\omega) - S(\omega)\mathbf{G}(\omega)]^H\mathbf{Q}^{-1}(\omega)[\mathbf{X}(\omega) - S(\omega)\mathbf{G}(\omega)]. \tag{20}$$

The goal of the proposed framework is thus to maximize the above likelihood, given the observations $\mathbf{X}(\omega)$, gain matrix $\mathbf{G}(\omega)$, and noise covariance matrix $\mathbf{Q}(\omega)$. Note that the gain matrix $\mathbf{G}(\omega)$ requires information about the location of the source. Hence, the optimization is usually solved through hypothesis testing. That is, hypotheses are made about the source location, which gives $\mathbf{G}(\omega)$. The likelihood is then evaluated. The hypothesis that results in the highest likelihood is determined to be the output of the SSL algorithm.

Instead of maximizing the likelihood in (18), we minimize the following negative log-likelihood:

$$\begin{aligned} J &= -\log p(\mathbf{X}|S,\mathbf{G},\mathbf{Q}) \\ &= -\int_\omega \left[\log\rho_\omega - \frac{J(\omega)}{2}\right] d\omega \\ &= \frac{1}{2}\int_\omega J(\omega)d\omega - \Theta \end{aligned} \tag{21}$$

where $\Theta = \int_\omega \log\rho_\omega d\omega$ is a constant. Since we assume the probabilities over the frequencies are independent of each other, we may minimize each $J(\omega)$ separately by varying the unknown variable $S(\omega)$. Given that $\mathbf{Q}^{-1}(\omega)$ is a Hermitian symmetric matrix $\mathbf{Q}^{-1}(\omega) = \mathbf{Q}^{-H}(\omega)$, we may take the derivative of $J(\omega)$ with respect to $S(\omega)$, and set it to zero to yield

$$\frac{\partial J(\omega)}{\partial S(\omega)} = -\mathbf{G}(\omega)^T\mathbf{Q}^{-T}(\omega)[\mathbf{X}(\omega) - S(\omega)\mathbf{G}(\omega)]^* = 0. \tag{22}$$

Therefore

$$S(\omega) = \frac{\mathbf{G}^H(\omega)\mathbf{Q}^{-1}(\omega)\mathbf{X}(\omega)}{\mathbf{G}^H(\omega)\mathbf{Q}^{-1}(\omega)\mathbf{G}(\omega)}. \tag{23}$$

Interestingly, (23) is identical to the MVDR filter described by (16). This relationship between the MVDR beamformer and the ML estimator was discovered earlier in [11].

Substituting the above $S(\omega)$ into (20), we can write

$$J(\omega) = J_1(\omega) - J_2(\omega) \tag{24}$$

where

$$J_1(\omega) = \mathbf{X}^H(\omega)\mathbf{Q}^{-1}(\omega)\mathbf{X}(\omega) \tag{25}$$

$$J_2(\omega) = \frac{[\mathbf{G}^H(\omega)\mathbf{Q}^{-1}(\omega)\mathbf{X}(\omega)]^H\mathbf{G}^H(\omega)\mathbf{Q}^{-1}(\omega)\mathbf{X}(\omega)}{\mathbf{G}^H(\omega)\mathbf{Q}^{-1}(\omega)\mathbf{G}(\omega)}. \tag{26}$$

Note that $J_1(\omega)$ is not related to the hypothesized locations during hypothesis testing. Therefore, the ML-based SSL algorithm shall maximize

$$J_2 = \int_\omega J_2(\omega)d\omega = \int_\omega \frac{[\mathbf{G}^H(\omega)\mathbf{Q}^{-1}(\omega)\mathbf{X}(\omega)]^H\mathbf{G}^H(\omega)\mathbf{Q}^{-1}(\omega)\mathbf{X}(\omega)}{\mathbf{G}^H(\omega)\mathbf{Q}^{-1}(\omega)\mathbf{G}(\omega)}d\omega. \tag{27}$$

Due to (23), we can rewrite J_2 as

$$J_2 = \int_\omega \frac{|S(\omega)|^2}{[\mathbf{G}^H(\omega)\mathbf{Q}^{-1}(\omega)\mathbf{G}(\omega)]^{-1}}d\omega. \tag{28}$$

The denominator $[\mathbf{G}^H(\omega)\mathbf{Q}^{-1}(\omega)\mathbf{G}(\omega)]^{-1}$ can be shown to be the residue noise power after MVDR beamforming [25]. Hence, the ML-based SSL algorithm is equivalent to forming multiple MVDR beamformers along multiple hypothesis directions and picking that output direction which results in the highest SNR.

IV. An Efficient SSL Algorithm for Distributed Meeting Applications

The above derived ML framework is very general. For instance, a similar ML SSL framework was presented in [12]. There, the goal was not only to estimate the location of the sound source, but also its directionality. A model similar to (3) was used, but the noise covariance matrix was assumed to be diagonal, $\mathbf{Q}(\omega) = \sigma\mathbf{I}$, where σ is independent of the microphone index and frequency. This led to a simplified objective function

$$J_2 = \int_\omega \left|\sum_{i=1}^{P}\alpha^*(\omega)X_i(\omega)e^{j\omega\tau_i}\right|^2 d\omega. \tag{29}$$

It is not difficult to verify that under these assumptions, (29) can be easily obtained from (27).

On the other hand, (27) cannot be directly applied to perform SSL in our current distributed meeting applications. In particular:

- the algorithm is too complex. If a full covariance matrix $\mathbf{Q}(\omega)$ is used, a $P\times P$ matrix inversion has to be conducted for each frequency bin, and the associated matrix multiplication (e.g., $\mathbf{G}^H(\omega)\mathbf{Q}^{-1}(\omega)\mathbf{G}(\omega)$) has to be conducted for each frequency bin and for each hypothesis source location;
- reverberation is not modeled in (27);
- for directional microphone arrays, the gain vector $\mathbf{G}(\omega)$ remains undetermined.

In the following, we will revise the noise model so that it can take reverberation into consideration. The $\mathbf{Q}(\omega)$ matrix will be diagonalized for fast computation. The gain vector will be explicitly estimated from the received signals and the noise model.

A. Reverberation

The reverberation of the room environment can be modeled as follows:

$$\mathbf{N}^r(\omega) = S(\omega)\mathbf{H}(\omega) \tag{30}$$

where $\mathbf{H}(\omega) = [H_1(\omega),\ldots,H_P(\omega)]^T$ is the room response function. We define the combined total noise as

$$\mathbf{N}^c(\omega) = \mathbf{N}^r(\omega) + \mathbf{N}(\omega) \tag{31}$$

and we assume the combined noise still follows a zero-mean, independent between frequencies, joint Gaussian distribution. The covariance matrix is

$$\begin{aligned}\mathbf{Q}^c(\omega) &= E\{\mathbf{N}^c(\omega)[\mathbf{N}^c(\omega)]^H\}\\ &= E\{\mathbf{N}(\omega)\mathbf{N}^H(\omega)\} + |S(\omega)|^2E\{\mathbf{H}(\omega)\mathbf{H}^H(\omega)\}\end{aligned} \tag{32}$$

where $E\{\cdot\}$ stands for expectation. Here we assume the noise and the reverberation are uncorrelated.

It should be noted that combining the reverberation term and the noise term to form a combined noise is not the only method to handle reverberation. In [28], Warsitz and Haeb-Umbarch included the room reflection function in the gain term $\mathbf{G}(\omega)$, and performed beamforming through an optimization algorithm that can obtain the combined $\mathbf{G}(\omega)$ directly. Nevertheless, their algorithm only computes the $\mathbf{G}(\omega)$ to optimize the beamformer's output SNR, which cannot be directly used to derive the actual sound source location $\mathbf{s}$ in (6).

The first term in (32) can be directly estimated from the silence periods of the acoustic signals

$$E(N_i(\omega)N_j^*(\omega)) = \lim_{K\to\infty}\frac{1}{K}\sum_{k=1}^{K}N_{ik}(\omega)N_{jk}^*(\omega) \tag{33}$$

where k is the index of audio frames that are silent. Note that the background noises received at different microphones may be correlated, such as the ones generated by computer fans in the room. If we believe the noises are reasonably independent at different microphones, we can simplify the first term of (32) further as a diagonal matrix

$$E\{\mathbf{N}(\omega)\mathbf{N}^H(\omega)\} = \mathrm{diag}(E\{|N_1(\omega)|^2\},\ldots,E\{|N_P(\omega)|^2\}). \tag{34}$$

The second term in (32) is related to reverberation. It is generally unknown. For efficient computation, we assume it is also a diagonal matrix

$$|S(\omega)|^2E\{\mathbf{H}(\omega)\mathbf{H}^H(\omega)\} \approx \mathrm{diag}(\lambda_1(\omega),\ldots,\lambda_P(\omega)) \tag{35}$$

with the i^{th} diagonal element equal to

$$\begin{aligned}\lambda_i(\omega) &= E\{|H_i(\omega)|^2|S(\omega)|^2\}\\ &\approx \gamma(|X_i(\omega)|^2 - E\{|N_i(\omega)|^2\})\end{aligned} \tag{36}$$

where $0 < \gamma < 1$ is an empirically determined parameter. Equation (36) assumes that the reverberation energy is a fraction of the difference between the total received signal energy and the environmental noise energy; the same assumption was used in [3], [15], and in (12). Note again that (35) is an approximation; reverberation signals received at different microphones are usually correlated, and the matrix should have nonzero off-diagonal elements. Unfortunately, it is generally difficult to estimate the actual reverberation signals or these off-diagonal elements. In addition, a nondiagonal noise covariance matrix would be very expensive to compute in practice.

As a result, we retain a diagonal covariance matrix for the combined noise

$$\mathbf{Q}^c(\omega) = \mathrm{diag}(\kappa_1(\omega), \ldots, \kappa_P(\omega)) \tag{37}$$

with the i^{th} diagonal element

$$\begin{aligned}\kappa_i(\omega) &= \lambda_i(\omega) + E\{|N_i(\omega)|^2\} \\ &= \gamma|X_i(\omega)|^2 + (1-\gamma)E\{|N_i(\omega)|^2\}.\end{aligned} \tag{38}$$

Equation (27) can thus be written as

$$J_2 = \int_\omega \frac{1}{\sum_{i=1}^{P} \frac{|\alpha_i(\omega)|^2}{\kappa_i(\omega)}} \left| \sum_{i=1}^{P} \frac{\alpha_i^*(\omega)}{\kappa_i(\omega)} X_i(\omega) e^{j\omega\tau_i} \right|^2 d\omega. \tag{39}$$

B. Estimating the Gain Factors

The gain factor $\alpha_i(\omega)$ can be accurately measured in some applications. For applications where it is unknown, we may assume it is a positive real number and estimate it as follows:

$$\begin{aligned}|\alpha_i(\omega)|^2|S(\omega)|^2 &\approx |X_i(\omega)|^2 - \kappa_i(\omega) \\ &= (1-\gamma)(|X_i(\omega)|^2 - E\{|N_i(\omega)|^2\})\end{aligned} \tag{40}$$

where both sides represent the power of the signal received at microphone i without the combined noise (noise and reverberation). Therefore

$$\alpha_i(\omega) = \frac{\sqrt{(1-\gamma)(|X_i(\omega)|^2 - E\{|N_i(\omega)|^2\})}}{|S(\omega)|}. \tag{41}$$

Inserting (41) into (39), we get

$$J_2 = \int_\omega \frac{\left| \sum_{i=1}^{P} \frac{X_i(\omega)e^{j\omega\tau_i}}{\kappa_i(\omega)} \sqrt{|X_i(\omega)|^2 - E\{|N_i(\omega)|^2\}} \right|^2}{\sum_{i=1}^{P} \frac{1}{\kappa_i(\omega)} (|X_i(\omega)|^2 - E\{|N_i(\omega)|^2\})} d\omega. \tag{42}$$

The computational cost of the above SSL algorithm is slightly higher than that of SRP-PHAT, but still manageable. In Section VI, we will demonstrate the superior performance of (42) under various noisy conditions.

C. Discussion

The ML-SSL algorithm proposed in (42) is closely related to existing SRP SSL algorithms in the literature. For instance, when the SNR is very high, we have $|X_i(\omega)|^2 \gg E\{|N_i(\omega)|^2\}$. Subsequently, $\kappa_i(\omega) = \gamma|X_i(\omega)|^2$, (42) thus becomes

$$\begin{aligned}J_2 &= \int_\omega \frac{\left| \sum_{i=1}^{P} \frac{|X_i(\omega)|}{\gamma|X_i(\omega)|^2} X_i(\omega) e^{j\omega\tau_i} \right|^2}{\sum_{i=1}^{P} \frac{|X_i(\omega)|^2}{\gamma|X_i(\omega)|^2}} d\omega \\ &= \frac{1}{\gamma P} \int \left| \sum_{i=1}^{P} \frac{X_i(\omega)e^{j\omega\tau_i}}{|X_i(\omega)|} \right|^2 d\omega\end{aligned} \tag{43}$$

which is equivalent to SRP-PHAT (10). Note that since the reverberation parameter γ is a constant factor of J_2, it does not affect the optimality of SRP-PHAT as long as the noise is very low.

The connection between the proposed ML-SSL algorithm and the ML algorithm in (11) may be not immediately evident. Recall that in their original derivation, Brandstein *et al.* [9] estimated the variance of the phase for a particular frequency as:

$$Var[\theta_i(\omega)] = \frac{E\{|N_i(\omega)|^2\}}{|X_i(\omega)|^2}. \tag{44}$$

If we ignore reverberation by setting $\gamma = 0$, and assume noise is relatively small compared to the signal (the same assumptions were made in [9]), then (42) can be written as

$$J_2 = \int_\omega \frac{\left| \sum_{i=1}^{P} \frac{e^{\theta_i(\omega)} e^{j\omega\tau_i}}{\frac{E\{|N_i(\omega)|^2\}}{|X_i(\omega)|^2}} \right|^2}{\sum_{i=1}^{P} \frac{|X_i(\omega)|^2}{E\{|N_i(\omega)|^2\}}} d\omega. \tag{45}$$

Therefore, the phase term of each microphone $e^{\theta_i(\omega)}e^{j\omega\tau_i}$ is indeed weighted by the inverse of the phase variance, as given by (44). Hence, the ML algorithm in (11) is conceptually similar to the proposed algorithm. On the other hand, the proposed ML-SSL algorithm differs from (11) by the presence of the additional frequency-dependent weighting (denominator in (45)). Furthermore, it has a more rigorous derivation, which demonstrates that it is a true ML algorithm for multiple microphones.

To summarize, Fig. 2 shows a relationship digram of the SSL algorithms mentioned in this paper. Note we use SRP-RUI to represent the algorithm proposed in [21] (13). The annotations on the links between algorithms indicate the conditions to simplify or convert one algorithm to the other. Note the proposed ML-SSL algorithm is the only one in this graph that is optimal in the ML sense for multiple microphones.

V. Enhanced MVDR Beamforming (eMVDR)

As presented in Section III, the MVDR algorithm, although derived from a very different perspective, is indeed identical to an intermediate step (23) during the derivation of the ML-SSL

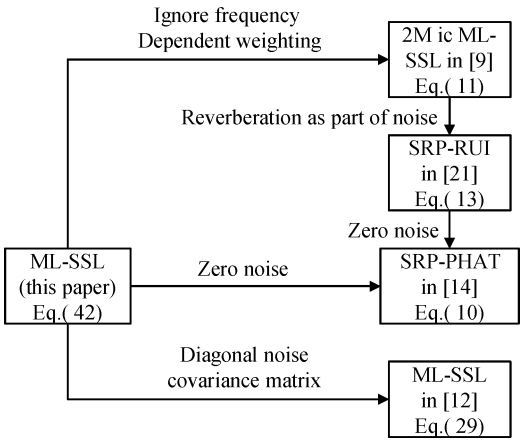


Fig. 2. Relationship diagram of the SSL algorithms mentioned in this paper.

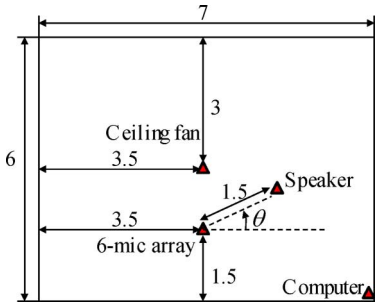


Fig. 3. Floor plan of the virtual room for synthetic experiments—distances shown in meters.

algorithm. Recent MVDR research has mostly focused on how to make MVDR robust to source location errors (such errors are usually caused by SSL). In this section, we propose an eMVDR approach that tries to address the problem of unknown directional patterns of microphones. It should be noted that existing robust MVDR algorithms can still be applied on top of our method to further improve performance.

Unlike in SSL, where reverberation can cause errors in the output source location, reverberation in beamforming is usually less of a concern in distributed meetings, because the reflected signal can still contain intelligible information. Therefore, in the following discussion, we ignore the reverberation term introduced during SSL (30), and use a noise covariance matrix directly estimated from the silence periods of the meeting

$$\mathbf{Q}(\omega) = E\{\mathbf{N}(\omega)\mathbf{N}^H(\omega)\}. \tag{46}$$

We start our discussion with (23) and (41).

From (41), it can be seen that $\alpha_i(\omega)$ can be estimated from the received signal and the noise model, though it is also related to the actual source energy. Fortunately, in MVDR it is the *relative* gains among the sensors that really matter in terms of beam shaping. Therefore, we define

$$\hat{\alpha}_i(\omega) = \frac{\alpha_i(\omega)}{\sum_{j=1,\ldots,P} \alpha_j(\omega)} \tag{47}$$

$$= \frac{\sqrt{(|X_i(\omega)|^2 - |N_i(\omega)|^2)}}{\sum_{j=1,\ldots,P} \sqrt{(|X_i(\omega)|^2 - |N_i(\omega)|^2)}}. \tag{48}$$

A new gain vector

$$\hat{\mathbf{G}}(\omega) = [\hat{\alpha}_1(\omega)e^{-j\omega\tau_1}, \ldots, \hat{\alpha}_P(\omega)e^{-j\omega\tau_P}]^T \tag{49}$$

is then inserted into (23) to perform MVDR.

It should be noted that by replacing actual gains with relative gains, we no longer compensate for the frequency response of the microphones. For example, if the microphones' average gain in a certain frequency is high, this will not be compensated for, and the final output signal will be stronger at that frequency. We believe this is not generally a problem, as most microphones have reasonably flat frequency responses. Furthermore, any equalization method used with a single mic could be similarly applied here, after beamforming.

VI. Experimental Results

We now present the results of our SSL and beamforming experiments. We run simulations on both synthetic signals and with extensive natural data.

A. SSL

We test the performance of the ML-SSL algorithm embodied in (42), on both synthetic and real-world datasets. The two benchmark algorithms we use to compare with the ML-SSL method are SRP-PHAT (10) and its improved version from [21] (13). Note that SRP-PHAT is a special case of SRP-RUI when $\gamma = 1.0$, while SRP-RUI is a special case of the ML-SSL algorithm when $\alpha_i(\omega) \equiv \alpha(\omega)$, $i = 1, \ldots, P$, and the frequency weightings are ignored.

1) Experiments on Synthetic Data: A virtual room with size $7 \times 6 \times 2.5$ meters is created, as shown in Fig. 3. A circular 6-microphone array is placed near the center of the room, at (3.5, 1.5, 1). The radius of the microphone array is 0.135 m. A speaker is talking at a distance of 1.5 m from the center of the microphone array, at an angle θ from the x-axis in Fig. 3. We introduce two noise sources in the scene. A ceiling fan is mounted in the middle of the room, at (3.5, 3, 2.5), and a computer is located in the corner, at (7, 0, 0.5). The wave signals from the speaker, the fan and the computer are all recordings from the real world. The reverberation effect of the room is added to all signals according to the image model [29].

The SSL algorithm performs hypothesis testing at 4° intervals in azimuth. The reported results are averaged over ten speaker locations uniformly distributed around the microphone array ($\theta = 0, 36^\circ, \ldots, 324^\circ$). At each location, the signal length is 30 s. The algorithm employs 40-ms windows spacing 20 ms apart. We sample 100 speech frames from each location and perform SSL on them. Table I reports the average accuracy, in terms of what portion of the SSL estimates (out of a total of $100 \times 10 = 1000$ frames) is within 2° and 10° of the ground truth angle. To assess the impact of reverberation on SSL performance, we synthesize rooms with 100 and 500 ms reverberation times, as seen in the upper and lower parts of Table I, respectively.

TABLE I
EXPERIMENTAL RESULTS OF SSL ACCURACY ON THE SYNTHETIC DATASET. CELLS WITH BOLD FONTS INDICATE BEST PERFORMANCE IN THE GROUP. (A) Reverberation = 100 ms. (B) Reverberation = 500 ms.

Input SNR	SRP-PHAT		$\gamma = 0.1$				$\gamma = 0.3$				$\gamma = 0.5$			
			SRP-RUI		ML-SSL		SRP-RUI		ML-SSL		SRP-RUI		ML-SSL	
	<2°	<10°	<2°	<10°	<2°	<10°	<2°	<10°	<2°	<10°	<2°	<10°	<2°	<10°
25 dB	97.6%	98.9%	97.2%	**99.1%**	**97.9%**	98.8%	96.8%	98.9%	**97.9%**	98.9%	96.2%	98.8%	97.8%	98.9%
20 dB	92.0%	93.6%	92.2%	**94.9%**	92.8%	94.7%	92.0%	94.4%	**93.0%**	**94.9%**	91.8%	94.3%	92.7%	94.6%
15 dB	89.0%	91.4%	90.8%	93.7%	**91.6%**	**93.9%**	90.2%	93.3%	91.5%	93.8%	89.5%	92.6%	91.2%	93.7%
10 dB	85.2%	88.8%	88.7%	91.4%	**89.0%**	**91.7%**	87.7%	90.7%	88.8%	90.9%	87.2%	90.2%	88.1%	90.4%
5 dB	76.1%	82.0%	86.1%	89.7%	**87.2%**	**90.3%**	82.7%	88.0%	85.9%	89.7%	80.7%	86.4%	85.2%	89.2%
0 dB	64.5%	71.1%	78.3%	85.7%	**81.2%**	**88.0%**	72.6%	80.5%	77.4%	84.0%	70.1%	77.3%	75.7%	82.9%

(a)

Input SNR	SRP-PHAT		$\gamma = 0.1$				$\gamma = 0.3$				$\gamma = 0.5$			
			SRP-RUI		ML-SSL		SRP-RUI		ML-SSL		SRP-RUI		ML-SSL	
	<2°	<10°	<2°	<10°	<2°	<10°	<2°	<10°	<2°	<10°	<2°	<10°	<2°	<10°
25 dB	**60.1%**	**79.2%**	53.8%	76.9%	60.0%	78.8%	54.1%	76.9%	59.8%	78.8%	54.3%	76.9%	59.9%	78.8%
20 dB	59.4%	78.4%	54.0%	77.0%	**60.3%**	**78.9%**	53.4%	76.8%	59.7%	78.7%	53.4%	77.1%	59.6%	78.6%
15 dB	60.3%	78.0%	55.2%	77.1%	**60.4%**	**78.8%**	53.4%	76.6%	60.1%	78.5%	53.1%	76.4%	59.6%	78.4%
10 dB	58.8%	77.0%	54.0%	76.5%	**59.8%**	77.1%	53.8%	76.1%	59.5%	77.6%	52.7%	75.8%	59.2%	**77.7%**
5 dB	56.3%	**75.5%**	53.1%	74.4%	**57.4%**	75.2%	52.5%	73.5%	57.2%	**75.5%**	52.3%	74.0%	57.1%	75.4%
0 dB	54.5%	74.4%	53.3%	73.9%	**56.2%**	74.4%	52.4%	74.9%	55.6%	74.8%	52.0%	74.5%	55.2%	**75.3%**

(b)

It can be observed from Table I that SRP-PHAT usually performs as well as ML-SSL when the input SNR is high (20 dB or above), but its performance drops significantly when the SNR becomes low. In most indoor (e.g., offices and meeting rooms) environments, the SNR is above 15 dB, which explains SRP-PHAT's satisfactory performance in practice. SRP-RUI is a very decent and practical SSL algorithm too. In low reverberation environment [Table I(a)], SRP-RUI has slightly worse performance than ML-SSL, and both algorithms significantly outperform SRP-PHAT in noisy cases. In high reverberation environments [Table I(b)], all three algorithms have a significant performance drop. ML-SSL still outperforms both SRP-PHAT and SRP-RUI, though by a small margin.

For the ML-SSL algorithm, the tunable parameter γ does seem to impact the final performance. This is particularly true when the reverberation is low. For instance, in Table I(a), when the reverberation is low (100 ms), when the input SNR is 0 dB, choosing $\gamma = 0.1$ results in much better performance than $\gamma = 0.5$. However, this gap is not significant when reverberation is high (Table I(b), 500 ms). Therefore, for practical applications, using a fixed γ ranging from 0.1 to 0.3 can usually result in satisfactory performance.

2) Experiments on Real-World Data: We next test the ML-SSL algorithm on 99 real-world meetings captured by the RoundTable device (Fig. 1). SSL is used in RoundTable to determine for which speaker the high-resolution video is to be provided. The main challenge of SSL for the RoundTable device is that the microphones are directional (in order to capture better audio) and they are arranged on a circle pointing in different directions. For microphones pointing away from the speaker, the estimated phase may be unreliable. In [21], the authors attempt to address the issue by selecting a subset of the microphones for SSL. In this paper, we use all the microphones, since ML-based SSL weights microphones differently based on their SNR automatically. We will compare our results with [21].

The meetings are each 4 min long, captured in about 50 different meeting rooms in order to test the robustness of the SSL algorithms in different environments. The noise levels of the rooms and the distances from the speakers to the devices vary significantly, causing the input SNR to range from 5 to 25 dB. The speaker locations of 6706 audio frames are labeled manually based on the corresponding face locations in the panoramic image. We report the results on the percentage of frames that are within 6° and 14° of the ground truth azimuth angle. This is slightly relaxed from the synthetic experiment but good enough for detecting speaker orientation in RoundTable.

The experimental results are shown in Table II. It can be seen that ML-SSL outperforms the other algorithms on this challenging dataset. The absolute accuracy improvement of ML-SSL over SRP-PHAT is about 2%. In terms of error rate, there is a 17% reduction. Fig. 4 shows a detailed accuracy plot (14° criterion) of the two algorithms on the 99 tested sequences. The dashed line is the equal accuracy line. It can be seen that ML-SSL outperforms SRP-PHAT in most sequences, and in some by a very significant margin.

The algorithm in [21], which selects a subset of microphones for SSL, is more efficient than SRP-PHAT; however, its performance is slightly worse. This shows that although there may be microphones which point in the opposite direction of the sound source and which may be very noisy, it may still be beneficial to include them for SSL.

TABLE II
EXPERIMENTAL RESULTS OF SSL ACCURACY ON THE REAL-WORLD DATASET. CELLS WITH BOLD FONTS INDICATE BEST PERFORMANCE IN THE GROUP

SRP-PHAT		Alg. in [21]		$\gamma = 0.2$				$\gamma = 0.5$			
				SRP-RUI		ML		SRP-RUI		ML	
<6°	<14°	<6°	<14°	<6°	<14°	<6°	<14°	<6°	<14°	<6°	<14°
81.73%	88.13%	80.55%	86.85%	83.06%	89.76%	**83.49%**	**90.13%**	82.76%	89.31%	83.03%	89.96%

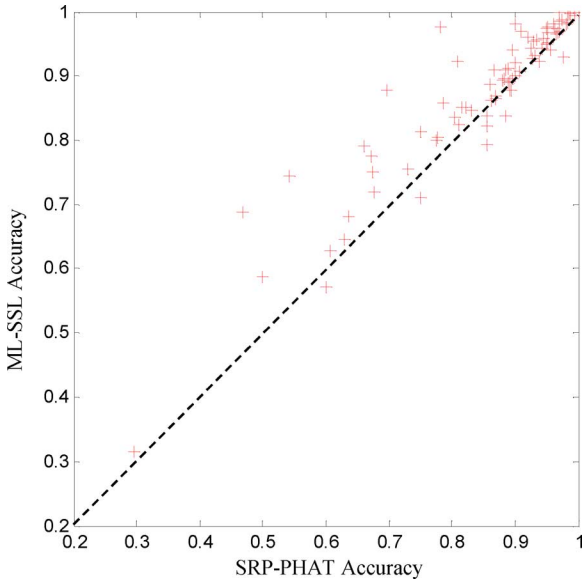


Fig. 4. Comparison of the SRP-PHAT accuracy and the ML-SSL accuracy. Each cross represents one real-world sequence. It can be seen that most of the crosses are above the diagonal dashed line, indicating that ML-SSL outperforms SRP-PHAT in most sequences.

B. eMVDR

In this section, we present some experimental results on beamforming using the circular array of directional microphones in the RoundTable device. Specifically, we compare the eMVDR beamformer to an MVDR algorithm without gain compensation as well as to a simple microphone selection scheme. The latter scheme chooses the microphone with the highest SNR as the estimate of the desired signal s. We have seen, based on extensive internal tests, that this scheme performs surprisingly well for the array in question. The SNRs of the output signals are compared against each other to evaluate performance.

We randomly picked ten audio sequences from the large dataset used above. We manually segmented the speech frames for the computation of the SNR. Table III summarizes the SNR results of the experiments. It can be seen that the eMVDR algorithm always outperforms traditional MVDR beamforming in terms of SNR. The average performance gain is 2.4 dB. The eMVDR beamformer also outperforms the best mic selection scheme by an average of 3.5 dB. It is interesting to note that there are two cases (E and I) where best mic selection does better than traditional MVDR beamforming and one (E) where it does better than eMVDR. Sequence I corresponds to a case with high SSL accuracy, which shows that not compensating for the directionality of the microphones can turn out to be expensive at times in terms of degrading the performance of the beamformer. Case E is a scenario where SSL accuracy is not as good as in I. This suggests that, in case E, one of the directional microphones might have been pointing directly at the source (meaning that the best mic selection scheme might have had a better estimate of the DOA of the source in case I). This may explain why, even after compensating for the gain pattern of the microphones, eMVDR still does slightly worse than best mic selection in case E. The performance loss in this case could be attributed to SSL accuracy.

TABLE III
COMPARISONS, BASED ON SNR (DB), OF BEST MIC SELECTION, MVDR AND *e*MVDR ON TEN AUDIO CLIPS

Clip ID	Best Mic	MVDR	eMVDR	SSL Accuracy
A	10.6	12.7	13.9	92.5%
B	19.8	21.5	25.5	95.5%
C	16.2	16.8	19.6	72.6%
D	22.6	24.2	25.2	98.3%
E	23.3	22.6	22.9	73.2%
F	18.8	21.8	24	93.9%
G	13.4	14.2	17.7	82.1%
H	20.2	21.1	23	45.1%
I	19.3	18.8	24.4	97.6%
J	14	14.9	16.4	54.4%
Avg.	17.8	**18.9**	**21.3**	80.5%

The results presented in Table III highlight several important points. First, they underline the importance of compensating for the gain pattern of directional microphones when using MVDR beamforming for speech enhancement. The enhancement of the proposed algorithm was 3.5 dB, compared to the 1.1-dB enhancement produced by traditional MVDR beamforming. Second, very much to our surprise, the best mic selection scheme does not seem to be such a bad algorithm after all. It has very low computational complexity and has performance comparable to that of traditional MVDR beamforming and similar to that of eMVDR (at least in the high SNR cases). However, we believe that the advantage of eMVDR beamforming over best mic selection comes from the fact that the average improvement in SNR can allow us to be more conservative in nonlinear post-processing operations. Since the post-processing usually distorts the signal when SNR is not high enough, a few dBs of improvement in SNR through eMVDR beamforming could actually result in better perceptual audio quality after post-processing.

VII. CONCLUSION

We have presented a ML framework for sound source localization and beamforming with microphone arrays. The main contribution of this paper is the novel adoption of an ML framework, leading to an efficient algorithm which works very well in practice. In particular, the proposed ML-SSL framework can

handle reverberant environments and unknown directional patterns of the microphones, and is linear in the number of microphones. Extensive experiments have demonstrated the superior performance of the proposed ML-SSL methods. The results on over 400 min of captured audio/video signals showed a 17% reduction in error rate, i.e., an increase in accuracy from 88.13% to 90.13%.

Close inspection of the remaining errors indicates a surprisingly high number of multi-source frames, which happens often during daily meetings [30]. Future work will include extending the current framework to multisource scenarios.

REFERENCES

[1] W. Wahlster, N. Reithinger, and A. Blocher, "Smartkom: Multimodal communication with a life-like character," in *Proc. Eurospeech*, Aalborg, Denmark, Sep. 2001.

[2] S. Basu, B. Clarkson, and A. Pentland, "Smart headphones: Enhancing auditory awareness through robust speech detection and source localization," in *Proc. IEEE ICASSP*, Salt Lake City, UT, 2001, vol. 5, pp. 3361–3364.

[3] H. Wang and P. Chu, "Voice source localization for automatic camera pointing system in videoconferencing," in *Proc. IEEE ICASSP*, New Paltz, NJ, Oct. 1997.

[4] R. Cutler, Y. Rui, A. Gupta, J. Cadiz, I. Tashev, L. W. He, A. Colburn, Z. Zhang, Z. Liu, and S. Silverbert, "Distributed meetings: A meeting capture and broadcasting system," in *Proc. ACM Conf. Multimedia*, 2002.

[5] M. Coen, "Design principles for intelligent environments," in *Proc. National Conf. Artificial Intelligence*, 1998.

[6] [Online]. Available: http://www.microsoft.com/presspass/presskits/uc/gallery.mspx

[7] C. Knapp and G. Carter, "The generalized correlation method for estimation of time delay," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-24, no. 4, pp. 320–327, 1976.

[8] L. J. Griffiths and C. W. Jim, "An alternative approach to linearly constrained adaptive beamforming," *IEEE Trans. Anttenas Propag.*, vol. AP-30, no. 1, pp. 27–34, Jan. 1982.

[9] M. Brandstein and H. Silverman, "A practical methodology for speech localization with microphone arrays," *Comput., Speech, Lang.*, vol. 11, no. 2, pp. 91–126, 1997.

[10] I. Ziskind and M. Wax, "Maximum likelihood localization of multiple sources by alternating projection," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 36, no. 10, pp. 1553–1560, Oct. 1988.

[11] K. Harmanci, J. Tabrikian, and J. L. Krolik, "Relationships between adaptive minimum variance beamforming and opitmal source localization," *IEEE Trans. Signal Process.*, vol. 48, no. 1, pp. 1–12, Jan. 2000.

[12] B. Mungamuru and P. Aarabi, "Enhanced sound localization," *IEEE Trans. Syst., Man, Cybern. B, Cybern.*, vol. 34, no. 3, pp. 1526–1540, Jun. 2004.

[13] X. Sheng and Y.-H. Hu, "Maximum likelihood multiple-source localization using acoustic energy measurements with wireless sensor networks," *IEEE Trans. Signal Process.*, vol. 53, no. 1, pp. 44–53, Jan. 2005.

[14] M. Brandstein and H. Silverman, "A robust method for speech signal time-delay estimation in reverberant rooms," in *Proc. ICASSP*, Munich, Germany, Apr. 1997.

[15] Y. Rui and D. Florêncio, "Time delay estimation in the presence of correlated noise and reverberation," in *Proc. ICASSP*, Montreal, QC, Canada, May 2004.

[16] J. Weng and K. Y. Guentchev, "Three-dimensional sound localization from a compact non-coplanar array of microphones using tree-based learning," *J. Acoust. Soc. Amer.*, vol. 110, no. 1, pp. 310–323, 2001.

[17] J. Kleban, "Combined Acoustic and Visual Processing for Video Conferencing Systems," Tech. Rep. Rutgers—The State University of New Jersey, New Brunswick, NJ, 2000.

[18] P. Georgiou, C. Kyriakakis, and P. Tsakalides, "Robust time delay estimation for sound source localization in noisy environments," in *Proc. WASPAA*, New Paltz, NY, Oct. 1997.

[19] T. Gustafsson, B. Rao, and M. Trivedi, "Source localization in reverberant environments: Performance bounds and ML estimation," in *Proc. ICASSP*, Salt Lake City, UT, May 2001.

[20] D. Li and S. Levinson, "Adaptive sound source localization by two microphones," in *Proc. Int. Conf. Robotics and Automation*, Washington, DC, 2002.

[21] Y. Rui, D. Florêncio, W. Lam, and J. Su, "Sound source localization for circular arrays of directional microphones," in *Proc. ICASSP*, Honolulu, HI, Apr. 2005.

[22] L. J. Griffiths and C. W. Jim, "An alternative approach to linearly constrained adaptive beamforming," *IEEE Trans. Anttenas Propag.*, vol. AP-30, no. 1, pp. 27–34, Jan. 1982.

[23] O. Hoshuyama, A. Sugiyama, and A. Hirano, "A robust adaptive beamformer for microphone arrays with a blocking matrix using constrained adaptive filters," *IEEE Trans. Signal Process.*, vol. 47, no. 10, pp. 2677–2684, Oct. 1999.

[24] [Online]. Available: http://bitwave.com.sg/products.php?current-page=computers

[25] H. Cox, R. M. Zeskind, and M. M. Owen, "Robust adaptive beamforming," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-35, no. 10, pp. 1365–1376, Oct. 1987.

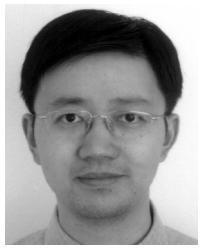
[26] B. D. Van Veen and K. M. Buckley, "Beamforming: A versatile approach to spatial filtering," *IEEE Signal Process. Mag.*, vol. 5, no. 2, pp. 4–24, Apr. 1988.

[27] A. El-Keyi, T. Kirubarajan, and A. Gershman, "Robust adaptive beamforming based on the kalman filter," *IEEE Trans. Signal Process.*, vol. 53, no. 8, pp. 3032–3041, Aug. 2005.

[28] E. Warsitz and R. Haeb-Umbach, "Acoustic filter-and-sum beamforming by adaptive principle component analysis," in *Proc. ICASSP*, Philadelphia, PA, Mar. 2005.

[29] J. B. Allen and D. A. Berkley, "Image method for efficiently simulating small-room acoustics," *J. Acoust. Soc. Amer.*, vol. 65, pp. 943–950, 1979.

[30] E. Shriberg, A. Stolcke, and D. Baron, "Observations on overlap: Findings and implications for automatic processing of multi-party conversation," in *Proc. Eurospeech*, Aaolborg, Denmark, Sep. 2001.



Cha Zhang (M'01) received the B.S. and M.S. degrees from Tsinghua University, Beijing, China, in 1998 and 2000, respectively, both in electronic engineering, and the Ph.D. degree in electrical and computer engineering from Carnegie Mellon University, Pittsburgh, PA, in 2004.

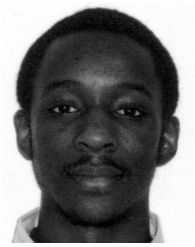
He is currently a Researcher in the Communication and Collaboration Systems Group at Microsoft Research, Redmond, WA. His current research focuses on applying various machine learning and computer vision techniques to multimedia applications, in particular, multimedia teleconferencing. He has published more than 30 technical papers and holds eight U.S. patents. He is co-author of *Light Field Sampling* (Princeton, NJ: Morgan and Claypool, 2006).

Dr. Zhang was the Publicity Chair for the International Packet Video Workshop in 2002 and the Program Co-Chair for the first Immersive Telecommunication Conference in 2007. He served as a Technical Program Committee member for numerous conferences, such as ACM Multimedia, CVPR, ICCV, ECCV, ICME, ICPR, and ICWL. He currently serves as an Associate Editor for the *Journal of Distance Education Technologies*. He won the Best Paper Award at ICME 2007.

Dinei Florêncio (SM'05) received the B.S. and M.S. degrees from the Universidade de Brasília, Brasília, Brazil, in 1983 and 1991, respectively, and the Ph.D. degree from the Georgia Institute of Technology, Atlanta, in 1996, all in electrical engineering.

Since 1999, he has been a Researcher with Microsoft Research, Redmond, WA. From 1996 to 1999, he was a member of the research staff at the David Sarnoff Research Center, Princeton, NJ. He has a passion for research that can have a real impact in products; his technologies have shipped in several Microsoft products and impacted the lives of millions of users. His current research interests include signal enhancement, 3-D video, and signal processing for collaboration and communication. He has published over 30 papers and holds 24 U.S. patents.

Dr. Florêncio received the 1998 Sarnoff Achievement Award. He will be a General Co-Chair of MMP '09.



Demba E. Ba received the B.S. (Hons.) degree in electrical engineering from the University of Maryland, College Park, in May 2004 and the M.S. degree in 2006 from the Massachusetts Institute of Technology, Cambridge, where he is currently pursuing the Ph.D. degree in electrical engineering.

His research interests lie in the areas of transform-based and multimedia signal processing, as well as statistical signal processing. He is interested in both theory and applications in the field.

Zhengyou Zhang (SM'97–F'05) received the B.S. degree in electronic engineering from the University of Zhejiang, Hangzhou, Zhejiang, China, in 1985, the M.S.degree in computer science from the University of Nancy, Nancy, France, in 1987, the Ph.D. degree in computer science from the University of Paris XI, Paris, France, in 1990, and the Ph.D. (Habilitation diriger des recherches) diploma from the University of Paris XI in 1994.

He joined Microsoft Research, Redmond, WA, in March 1998, where he is a Principal Researcher and manages the human-computer interaction and multimodal collaboration group. He has previously with the French National Institute for Research in Computer Science and Control (INRIA) for 11 years, where he was a Senior Research Scientist since 1991. During 1996–1997, he spent a one-year sabbatical as an Invited Researcher at Advanced Telecommunications Research Institute International (ATR), Kyoto, Japan. He holds more than 50 U.S. patents and has about 40 patents pending. He also holds a few Japanese patents for his inventions during his sabbatical at ATR. He has published over 160 papers in refereed international journals and conferences, has edited three special issues, and has co-authored three books: *3D Dynamic Scene Analysis: A Stereo Based Approach* (Berlin/Heidelberg: Springer, 1992); *Epipolar Geometry in Stereo, Motion and Object Recognition* (Norwell, MA: Kluwer, 1996); and (in Chinese) *Computer Vision: Theory and Applications* (Beijing, China: Chinese Academy of Sciences: 1998 and 2003).

Dr. Zhang has been a member of the IEEE Computer Society Fellows Committee since 2005, Chair of the IEEE Technical Committee on Autonomous Mental Development, and a member of the IEEE Technical Committee on Multimedia Signal Processing. He is currently an Associate Editor of several international journals, including the IEEE TRANSACTIONS ON MULTIMEDIA, the *International Journal of Computer Vision*, the *International Journal of Pattern Recognition and Artificial Intelligence*, and the *Machine Vision and Applications* journal. He served on the Editorial Board of the IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE from 2000 to 2004, among others. He has been on the organization or program committees for numerous international conferences, and was a Program Co-Chair of the Asian Conference on Computer Vision (ACCV2004), January 2004, Jeju Island, Korea, a Technical Co-Chair of the International Workshop on Multimedia Signal Processing (MMSP06), October 3–6, 2006, Victoria, BC, Canada, and a Program Co-Chair of the International Workshop on Motion and Video Computing (WMVC07), February 23–24, 2007, Austin, TX.